AN AUTOMATIC METHOD OF CLASSIFYING MOLECULES

FIELD OF THE INVENTION

The present invention relates to an automatic method and system for the identification of biological families of molecules, in particular proteins, and a method for creating an hierarchical organization within those families.

BACKGROUND OF THE INVENTION

In recent years we have been witnessing a constant flow of new biological data. Large scale sequencing projects throughout the world turn out new sequences and create new challenges for investigators. Many sequences that are added to the databases are unannotated and await analysis. Currently, 13 complete genomes (of yeast, E. coli other bacteria and several archae) are available. Between 35%-50% of their proteins have an unknown function (Penmisi 1997, Doolittle 1998). In the absence of structural data, analysis necessarily starts by investigating the sequence proper. Sequence analysis considers the basic properties of individual amino acids, as well as their combination. The most effective analyses compare the sequence under study with the whole database, in search for close relatives. Properties of a new protein sequence are extrapolated from those of its neighbors. Since the early 70's, algorithms were developed for the purpose of comparing protein sequences (Needleman & Wunsch 1970, Smith & Waterman 1981, Lipman & Pearson 1985, Altschul et al. 1990).

It is generally accepted that two sequences with over 30% identity along much of the seequences, are very likely to have the same fold [Sander & Schneider 1991, Flores et al. 1993, Hilbert et al. 1993]. Proteins of the same fold usually have similar biological functions (with the exception of convergent evolution in which the same fold is shared by non homologous proteins). Nevertheless, one encounters many cases of high similarity both in fold and function, despite a low sequence similarity (Muzzin 1993, Pearson 1997). Such instances are, unfortunately, often missed by current search methods.

Detecting homology may often help in determining the function of new proteins. By definition, homologous proteins have evolved from the same ancestor protein. The degree of conservation varies among protein families. However, homologous proteins almost always have the same fold (Pearson 1996). Homology is, by definition, a transitive relation: If A is homologous to B, and B is homologous to C, then A is homologous to C. This simple observation can be very effective in discovering homology. However, when applied simple-mindedly, this observation leads to many pitfalls.

Though the common evolutionary origin of two proteins is almost never directly observed, we can deduce homology among proteins, with a high statistical confidence, given that the sequence similarity is significant. This is particularly useful in the so called "twilight zone" (Doolittle 1992), where sequences are identical to, say, 10-25%. Transitivity can be used to detect related proteins, beyond the power of a direct search.

Though transitivity is an attractive concept, some perils of transitivity should be taken into account as well. It should be clear that similarity is not transitive, and that it does not necessarily entail homology. Significant similarities can be used to infer homology, with a level of confidence that depends on the statistical significance (see Pearson 1996).Therefore similarity should be carefully used in attempting to deduce homology. The statistical significance level of the similarity should take into account the level of evolutionary divergence within the family, in order to deduce homology reliably.

Multidomain proteins make the deduction of homology particularly difficult. If protein 1 contains domains A and B, protein 2 contains domains B and C, protein 3 contains domains C and D, then should proteins 1 and 3 be considered homologous? This simple example indicates the inadequacy of single-linkage clustering for the purpose of identifying protein families within the sequence space. Expert biologists can distinguish significant from insignificant similarities. However, the sheer size of current databases rules out an exhaustive manual computation of homologies. Our goal was to develop an automatic method for classification of protein sequences based on sequence similarity, through the detection of groups of homologous proteins (clusters).

and high l vel structures (groups of related clusters) within the sequence space. Such organization would reveal relationships among protein families and yield deeper insights into the nature of newly discovered sequences.

SUMMARY OF THE INVENTION

It is an object of the present invention to offer a method for hierarchical organization of protein sequences in a database according to their biological functions, by using restricted transitivity for disclosing the similarity among proteins (splitting the space of all protein sequences into connected components or clusters wherein every two members are either directly or transitively related.)

It is another object of the present invention to identify many biological families. By varying the threshold of statistical significance, finer sub- families that make up known families of proteins are being discovered. Likewise, this procedure exposes linkages between distinct protein families. Broadly speaking, protein families turn out to be connected in two distinct ways: (i) Through multi domain proteins, each of which is associated with a distinct protein family, or (ii) Through proteins much of whose sequence is shared by the two families. The latter may be considered as linkers or ancestor proteins. Consequently, many interesting relations between protein families are revealed and hierarchical organization within protein families suggest themselves.

It is another object of the present invention to provide a new metric on the space of all protein sequences, which is more sensitive than the existing measures. Global self organization of all known protein sequences reveals inherent bilogical signatures. *JMB 268, 539-556* for analyzing new sequences (which are not on the SWISSPROT database) and defining them according to the hierarchical organization.

The present invention provides an automatic method of classifying molecules having similar biologic function comprising the steps of: a) creating a hierarchical organization of said molecules in a database, wherein groups of clusters are identified using local consideration resulting in related clusters; b) determining the position of a selected molecule based on the hierarchical organization of step a, whereby selected molecules of similar biologic function are classified.

The present invention further provides an automatic method of classifying molecules having similar biologic function comprising the steps of:

a) creating a hierarchical organization of the molecules in a database comprising the steps of:

i) calculating pairwise similarities between said molecules in said database by combining at least one standard measure of similarity, resulting in a first set of expectation values of similarity;

ii) analyzing said first set of expectation values of similarity so as to obtain a second set of expectation values of similarity, wherein said molecules of the second set have a high degree of similarity;

iii) merging said resulting molecules of step ii) so as to form clusters, wherein only molecules with an expectation value below a first restricted threshold are merged;

iv) identifying groups of clusters using local consideration resulting in related clusters;

v) determining the relationship between related clusters in said groups;

vi) analyzing said groups of the related clusters of step v), thereby creating a hierarchical organization of said molecules;

b) determining the position of a selected molecule based on the hierarchical organization of step a, comprising the steps of:

i identifying pairwise similarities between said selected molecule to the molecules in said database by combining at least one standard measure of similarity, resulting in a third set of expectation values of similarity;

ii identifying geometric averaging of said selected molecule to each of said resulting clusters in said hierarchy, resulting in a forth set of expectation values of similarity;

iii identifying related clusters from said hierarchy of step a, having a geometric averaging with said selected molecule below a second threshold, thereby classifying molecules having similar biologic function.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is herein described, by way in example only, with reference to the accompanying drawings. In the drawings:

Figures 1A-1C are graphic representations of the average distribution of e-values according to three main algorithms for sequence comparison, BLAST, FASTA and SW respectively, according to an embodiment of the invention;

Figure 2 is a schematic tree of the hirarchical organization of small G-protein family;

Figure 3 is a schematic tree of the hirarchical organization of ATP-binding transporter family;

Figure 4 is a schematic tree of the super- family of motor proteins;

Figure 5 is a schematic tree of the hirarchical organization of proteins that are involved in the biosynthesis of complex sugars;

Figure 6 is a schematic alignment of two protein sequences sw:acsa_acexy and sw:ythLrhoso;

Figure 7 is a schematic tree of the hirarchical organization of methylases and the methyltransferases;.

Figure 8 is a schematic alignment of two protein sequences sw:hmt1_yeast and sw:y912_haein;

Figure 9 is a table showing PROSITE families for which performance was below 50% true positives;

Figure 10 is a table showing correlation of releases according to an embodiment of the invention;

Figure 11A is a table showing clusters possibly related to cluster 4 of the immunoglobulin V region, according to an embodiment of the invention;

Figure 11B is a table showing clusters possibly related to cluster 5 of the immunoglobulins and major histocompatibility complex, according to an embodiment of the invention;

Figure 12 is a schematic illustration tracing the formation of clusters;

Figure 13 is a schematic illustartion showing the association of groups according to an embodiment of the invention;

Figure 14 is a schematic illustration of the clustering alsorithm according to an embodiment of the invention;

Figure 15 is a table showing the largest clusters at the lowest confidence level 10^{-0};

Figure 16 is a graphic representation of the correlation of BLAST e-values and SW e-values;

Figure 17 is a schematic representation of the hierarchial organization of the small G-protein family;

Figure 18 is a schematic map of the immunoglobulin superfamily according to an ambodimant of the invention;

Figure 19 is a table showing the distribution of cluster sizes at various confidence levels;

Figure 20 is a table showing performance evaluation;

Figure 21 is a table showing the Largest clusters for which there was no corresponding family in PROSITE 13 or Pfam 1.0; and

Figure 22 is a table showing clusters belonging to the immunoglobulin superfamily.

DETAILED DESCRIPTION OF THE INVENTION

The present invention overcomes the drawbacks of the prior art by using transitivity in a controlled way (restricted transitivity) enabling organization of all the proteins in the database in a hierarchical organization.

The present invention applied the following methods for comparing sequences against the databases: the Smith Waterman (SW), as disclosed in Smith, T. F. & Waterman, M. S. (1981), Comparison of Biosequenoes, Adv. *in Appl.* Math. 2, 482-489, the FASTA, as disclosed in Lipman, D. J. & Pearson, W. R. (1985), Rapid and sensitive protein similarity, Science 227, 1435-1441 and the BLAST, as described in Altschul, S. F., Carrol, R. J. & Lipman, D. J. (1990), Basic local alignment search tool, *J. Mol. Biol.* 215, 403-410. These papers are incorporated by reference in their entirety herein. The results are analysis at different thresholds. The analysis starts at a very high and restricted threshold, where the proteins are organized in small clusters. When the threshold is increased in a step wise manner the clusters merge to give larger clusters, subfamilies and finally families. When a protein belongs to a certain cluster it influences and is influenced by the other members of the clusters. Thus, the question whether a certain cluster will merge with another cluster depends on the similarity of all of the protein that are in the cluster. There could be a situation in which cluster A contains a protein showing similarity to cluster B and therefore tends to merge with cluster B, but the other members of cluster A which do not show the same degree of similarity will prevent this merge.

By performing this procedure at varying thresholds, in a stepwise manner a hierarchical organization of the connected components is obtained, and thus of all known proteins. As disclosed herein, the analysis of the invention starts from a very conservative classification, based on highly significant similarities, that consists of many classes. Subsequently, classes are merged to account for less significant similarities. Merging is performed via a two phase algorithm. First, the method identifies groups of possibly related clusters using local considerations, as will be described below. Subsequently, a global test is applied to identify nuclei of strong relationships within these groups of clusters, and clusters are merged accordingly. This process is repeated

at varying levels of statistical significance, obtaining a hierarchical organization of all proteins.

The resulting classification splits the protein space into well defined groups of proteins, which are closely correlated with natural biological families and superfamilies. Different indices of validity were tested to assess the quality of the classification. When compared with the protein families in PROSITE and Pfam databases, the classification has detected between 64.8% and 88.5% of the proteins in these families, as well as many new clusters which do not match any protein family in these databases. The hierarchical organization reveals finer subfamilies that make up known families of proteins as well as many novel relations between protein families.

The method of the present invention is carried out by software or software means (data) executable on computing (data processing) means, such as a computer (PC) or similar data processor, microprocessors, embedded processors, microcomputers, microcontrollers, etc. These computing means are capable of performing processes typically algorithms, performing the method of the present invention.

The procedure of the present invention was performed for the SWISSPROT database. As disclosed in Bairoch, A. & Boeckman, B. (1992) The SWISSPROT protein sequence data bank. (*Nucl. Acids Res. 20, 2019-2022).*

The procedure starts from creating a neighbors list for each sequence in the swissprot database in each of the following methods: the Smith Waterman (SW), the FASTA and the BLAST. A numerical normalization is applied first to all methods, so they are all on comparable scales. Then, only statistical significant similarities are maintained in these lists. These lists induce agraph to obtain a hierarchical organization of all of the proteins that in the database. The vertices of this graph are the protein sequences. Edges between the vertices are weighted with weights that reflect the distance or dissimilarity between the corresponding sequences. i.e. high similarity translates to a small weight (or distance). To compute the weight of the directed edge from A to B, one compares A against all sequences in the swissprot database, and obtains the distribution of its score. The weight is taken as the expectation value as disclosed in Atschul, S. F., Boguski, M. S., Gish, W. G. &

Wooton, J. C. (1994) Issues in searching molecular sequence databases. *Nature Genetics 6, 119-129* of the similarity score between A and B, based on this distribution. This is a statistical estimate for the number of occurrences of the appropriate score at a random setup, assuming the existing amino acid composition. When the similarity score is statistically insignificant, the corresponding edge is discarded. In other words. an edge among sequence A and B indicates that the corresponding proteins are likely to be related. Finally, the weight of an edge is defined as the minimum associated to it by any of the three methods, to capture the apparently strongest relation. When all edges of weight are eliminated below a certain significance threshold, the graph splits into connected components. These automatically induced sets of proteins that are closely correlated with natural biological families. By performing this procedure at varying thresholds, in a stepwise manner a hierarchical organization of the connected components is being obtained, and thus of all known proteins.

This graph encodes much of the fundamental properties of the sequence space. It reflects the idea that interesting homologies among proteins can be deduced by transitivity. The transitive closure of the similarity relation among the proteins shown by the graph, splits the space of all protein sequences into connected components or clusters. These are proper subsets of the whole database wherein every two members are either directly or transitively related. These sets are maximal in this respect and cannot be expanded. Thus they offer a self-organized classification of all protein sequences in the database. These connected components can be expected to correlate with known biological families. The results suggest that these connected components are indeed very informative and robust.

An attention should be drawn that this is a directed graph. so it is not necessarily symmetric. Specifically, it may (and does) happen that there is an edge from protein A to protein B, but none in the reverse direction. Furthermore, even if both edges exist, their weights may differ. Therefore, our notion of a component is that of a *strong* connected component. The partition into strongly connected components is thus more refined then the partition into connected components. This analysis can be performed at different thresholds, or confidence levels, to obtain an

hierarchical organization. Several connected components of a given threshold may fuse together at a more permissive threshold. The analysis starts at the 10^{-100} threshold. Subsequent runs are carried out for 10^{-95},10^{-90}, ...10^{-0} = 1. Almost all of the clusters that were found are meaningful. Some correspond to well known families, but many others correspond to less studied families. There are clusters that consist exclusively of unknown proteins or hypothetical proteins.

Level	*Over 100*	*51-100*	*21-50*	*11-20*	*6-10*	*2-5*	*1*	*Total no.*
10^{-100}	8	18	90	234	528	3727	29870	34475
10^{-95}	8	19	101	238	536	3806	29078	33786
10^{-90}	8	20	112	254	546	3866	28212	33018
10^{-85}	8	23	119	261	565	3997	27178	32151
10^{-80}	8	25	136	258	594	4060	26127	31208
10^{-75}	9	33	134	268	619	4113	25030	30206
10^{-70}	11	35	140	285	645	4117	23912	29145
10^{-65}	12	35	156	301	650	4154	22872	28180
10^{-60}	15	36	166	311	671	4134	21721	27054
10^{-55}	16	42	176	321	670	4143	20584	25952
10^{-50}	17	50	180	349	663	4140	19386	24785
10^{-43}	21	52	190	359	673	4109	18174	23578
10^{-40}	26	55	192	369	690	4034	17009	22375
10^{-35}	30	55	199	374	718	4012	15656	21044
10^{-30}	33	50	208	382	738	3903	14268	19582
10^{-25}	36	54	224	375	719	3725	12963	18096
10^{-20}	36	60	237	375	701	3536	11593	16538
10^{-15}	37	57	230	378	685	3222	10221	14830
10^{-10}	35	53	227	349	655	2855	8708	12882
10^{-3}	24	41	191	282	528	2352	6852	10263
10^{-0}	1	0	33	88	207	1292	4639	6260

Table 1:Distribution of clusters by their size at each confidence level. Table 1 shows the distribution of clusters by their size at the different confidence levels. At each level, the universe of all proteins splits into connected components (clusters). These clusters become larger and coarser with the decrease of confidence levels. Consequently, the number of isolated proteins (clusters of size 1) decrease. As confidence level decrease to 10^{-0} There is a sharp decline in the number of midsize clusters. Chance similarities tend to blur the picture, and cause an "avalanche", where (possibly unrelated) many families are joined to few giant clusters.

Reference is now made to Figs. 1A-1C. These figures are based on the following main algorithms for measuring similarity between protein sequences : the BLAST, the FASTA and the SW respectively. These methods are in daily use by biologists, for comparing sequences against the databases. Though SW tends to give the best results, it is not uncommon that FASTA or BLAST are more informative as disclosed in Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Science 4, 1145-1160* therefore the results of the three systems are being incorporated in order to achieve maximum sensitivity. For example SW is being used with the BLOSUM50 scoring matrix while FASTA and BLAST are being used with the BLOSUM62 scoring matrix in order to achieve better identification of remote homologous. Another example is when searches are strongly biased because the amino acid composition of the query sequence differs markedly from the overall average composition. A case in point is the effects of low complexity segments within the sequence as disclosed in Altschul, S. F., Boguski, M. S., Gish, W. G.& Wooton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genetics 6, 119-129.* Therefore, the results of BLAST (which uses the SEG program) as disclosed in Wootton, J. C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem. 17, 149-163* in amino acid are being used for excluding low complexity segments from the query sequence.

In order to compare the scores between the three methods one may pick any protein, carry out an exhaustive comparison against the whole database and consider

the highest score in each of the methods, then plot these values and compare two methods at a time. These scores show a very strong linear relation in log-log scale, therefore introducing a usually small multiplicative factor which per each protein and per method would scale the three methods to a single reference line.

The differences between FASTA and SW are mostly due to the different scoring matrices that are being used and can be corrected by multiplying the original score by the relative entropy of the two matrices as described in Altschul, S. F. (1991) Amino acid substitution matrices from an information theoretic perspective. *JMB 219, 555-565.* The difference between SW and BLAST may be due to approximations in estimating the parameters λ and K as described in S. Karlin & S. F. Adtsdiul. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *PNAS 87, 2264-2268.* The underlying assumption in calculating these parameters is that the amino acid composition of the query sequence is close to the overall distribution. This assumption often fails, e.g. for low complexity segments. Moreover, these parameters are based on first order statistics of the sequence, the scoring matrix and the database. The corrections that are required to match SW and BLAST may be due to inaccurate approximations of the estimated parameters, or due to higher order statistics of the sequence.

It is very difficult to set a clear dividing line between true homologies and chance similarities. Expectation values below 10^{-3} can be safely considered significant and those above 10 reflect almost pure chance similarity. However, the range within is difficult to characterize, and truly related proteins may have expectation values around 1. An overly strict threshold will miss important similarities within the twilight zone, whereas an excessively liberal criterion will create many false connections. The exact threshold for each method was set to best discern among related and unrelated proteins. In the present invention it is based on the overall distribution of distances over the entire protein space, as given by each of the three methods.

This is illustrated in Figs. 1A - 1C, which show the average distribution of expectation values over the entire SWISSPROT database, for BLAST, FASTA and SW respectively. The graphs in Figs. 1A-1C naturally suggest a threshold for each method. The distribution drawn a log-log scale is nearly linear, at low expectation

values, but starts a rapid increase at a certain value. This value is set to be the threshold. The thresholds for SW, FASTA and BLAST are set at 0.1, 0.1 and 10^{-3} respectively. An edge from vertex A to vertex B is maintained only if a significant score is obtained on comparing the corresponding proteins. Namely, if either SW or FASTA yield an expectation value ≤ 0.1 or BLAST's expectation value is $\leq 10^{-3}$.

A major difference between BLAST and SW/FASTA is that BLAST charges no gap penalties. Consequently, BLAST tends to overestimate the statistical significance of alignments. The method of the present invention counters this behavior of BLAST by selecting the edges asymmetrically as shown in figs. 1A-1C. While this property may help BLAST reveal significant similarities that the other methods miss (as described in Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Science 4, 145-1160)* it could also bring to highly fragmentary alignments that cannot be considered biologically meaningful. Therefore, the method of the present invention ignores those BLAST scores that come from a large number of HSPs (high scoring pairs), where the MSP (maximal segment pair) is insignificant.

Finally, even if the comparisons between proteins A and B fail to satisfy the previous criteria, the edge from A to *B* is maintained when all three methods yield an expectation value ≤ 1.

Reference is now made to Figs. 2 and 3. These figures are examples of hierarchical organization within known families. This organization is based on the information extracted while moving across the different levels of the tree which means scanning the hierarchy over all levels. Fig. 2 describes the small G-protein/Ras super family. This family of proteins is composed of the sub-families: ras, rab, ran, rho, ral, and smaller sub-families. Fig. 2 depicts the relations within this family, based on the hierarchical organization obtained by the method of the present invention. A total of 366 proteins, all belonging to the small G-protein super-family, are presented. Small clusters, which correspond to subfamilies, are formed at the high levels of confidences, and fuse to larger clusters, when the threshold is raised. At the low level of confidence of 10^{-10}, this family merges with the ADP ribosylation factors family and guanine nucleotide-binding proteins, all of which are GTP-binding proteins, to form one cluster.

Fig. 3 describes the ATP-binding transporters family. Transporters are membranous elements which provide the mechanism by which components cross the lipid bilayer within and from the environment of cell compartments. The large variety of components, environmental conditions and organisms make this super-family very complex and rich. This diverged family is another example for which an hierarchical organization is proposed. The sub-classification distinguishes between amino-acids transporters, oligopeptide transporters, metal transporters, multi drug resistance proteins, and many more subgroups. Out of 296 proteins presented here, 75 are hypothetical transporters which can be classified based on this organization, according to their position in the tree.

Reference is now made to Figs. 4,5 and 7. These figures demonstrate how transitivity can be used to verify the relation between different, but functionally related, protein families and how the connections are created when moving from one confidence level to the next level.

Fig. 4 describes the hierarchy of the super-family of motor proteins. Each circle indicated by the numeral 100, represents a connected component at threshold 10^{-35}. Circles' radii are proportionate to the component's size. The component's size appears next to the corresponding circle. The drawn edges appeared upon lowering the threshold to 10^{-30}. The letters A-D indicate the order of transitivity indicated by the vertical line 101. Each cluster is referred to by its order of transitivity, A-D, and its position from left to right. Clusters which consist solely of hypothetical proteins are indicated by a second superimposed dashed circle 102 .

Fig. 4 demonstrates how, in some cases, the connection between functionally related proteins is revealed only through the connection with hypothetical proteins. One such example is the connection between the myosins and the kinesins. The isolated sets (at the level of 10^{-35}) of kinesin and kinesin-like proteins (sets C1, D1, D2), myosins (set A1), axoneme-associated proteins (set C2) , and trichohyalin (set B2), were grouped together, in some cases via connections with hypothetical proteins (sets B1, C3), to form a super family of motor proteins (at the level of 10^{-30}) with a total of 120 proteins. All proteins share an elongated structure, and energy dependent motor activity and are expressed throughout the evolutionary tree. They do vary in

their directionality of action, tissue specificity and their highly diverged biological contexts.

Similarly Fig. 5 describes the hierarchy of proteins involved in the biosynthesis of complex sugars, where each circle represent a connected component at threshold 10^{-10}, and the drawn edges appeared upon raising the threshold to 10^{-5}. Indeed, a relation between those families is established thanks to the raising of the confidence level from 10^{-10} to 10^{-5}. As in the previous example, the connection is established via hypothetical proteins, based on weak alignments. However, the basic biological feature which characterizes all these proteins, makes the connections inevitable.

Fig. 6 describes the alignment of the sequences sw:acsa-acexy and sw:yth1-rhosom, (which correspond to sets A1 and B2 respectively in Fig. 5). Fig. 6 shows 24% identity and this demonstrates the weak alignment achieved when comparing protein sequences, where there are similar biological features connecting the two sequences. In contrast, Fig. 5 shows the connection of the sequences classified to sets A1 and B2. Thus it's possible to identify that these two sequences belong to the same biological family. This connection can only be achieved by the method of the present invention as was shown in Fig. 5.

Similarly Fig. 7 describes the hierarchy of proteins involved in the methylase, methyltransferase family where each circle represents a connected component at threshold 10^{-10}, and the drawn edges appeared upon raising the threshold to 10^{-5}. This family is another example for the importance of hypothetical proteins as linkage proteins. Through such links a natural connection between related biological families is established, as demonstrated for methylasels and methyltransferases.

A total of 80 proteins in 28 isolated sets (at the level of 10^{-10}) were connected to one cluster at the level of 10^{-5}. The Y-axes of the graph represents 11 orders of transitivity (labeled A-K). At the bottom end (A1, B1, B2,C1) as well as at the top end (J2) methylases and methyltransieraser are common. Many hypothetical proteins are scattered within these clusters. Some clusters contain exclusively hypothetical proteins (E2, H1, I2), The connection between the two ends of this graph is made through such clusters (H1,I2).

Fig. 8 describes the alignment of sequences sw:hmt1-yeast and sw:y912-haein that correspond to sets G1 and H1 respectively, in Fig. 7. Fig. 8 shows 30% identity which demonstrates the weak alignment achieved when comparing protein sequences, where there are similar biological features connecting the two sequences. In contrast Fig. 7 shows the connection of the sequences classified to sets G1 and H1. The connections are based on very sparse pairwise alignments as compared with Fig. 8, and thus raise a reasonable doubt on the biological significance. Yet, proteins at the two ends of the graph of Fig. 7 exhibit close biological function, therefore verify the validity of these connections.

EXPERIMENTAL DETAILS SECTION

Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space

Motif and domain based analyses: Most of these studies yielded databases of protein motifs and domains, which have become an important tool in the analysis of newly discovered protein sequences. Among these are PROSITE (Bairoch et al. 1997), PRINTS (Attwood et al. 19981, Blocks, HenikoffL& Henikoff 1991), Pfam (Sonnbammer et al. 1997), ProDom (Sonnhammer & Kahn 1994), and Domo (Gracy & Argos 1998). The manually defined patterns in PROSITE have served as an excellent seed for several such studies.

There are several aspects in which these studies differ from each other. Some are based on manual or semi-manual procedures (e.g,. PROSITE, PRINTS), others are semi-automatic (Pfam) and the rest are fully automatic (e.g. ProDom, Blocks, Domo). Some focus on short motifs (PROSITE, PPINTS, Blocks) while others seek whole domains and try to infer domain boundaries (Pfam, ProDom, Domo). Most databases also give the domain/motif structure of proteins. Two databases make use of transitivity to enhance sensitivity. Prodom applies the transitive closure of high scoring segments pairs obtained by BLAST (only when the common segments overlap, above a minimum overlap parameter). In the Pfam database, the construction

of new families starts form a model (HMM) derived from multiple alignment of related proteins, which is being improved iteratively by searching for more related sequences in the database. These are iteratively incorporated into the model, till convergence. In each iteration the resulting alignment is traced manually, to avoid misalignments.

Protein based analysis: These studies can be divided into two categories. The first category includes methods that employ alternative representations of protein sequences, e.g. their di peptide composition (van Heel et al. 1991, Ferran et al. 1994) or combination of compositional properties and other physical/chemical properties (Hobohm and Sander 1995). The alternative representations induce measures of similarity/dissimilarity between complete protein sequences. These measures can in turn, be used (van Heel et al. 1, Ferran et al. 1994) to classify the sequences into (usually a fixed number of) clusters. The second category includes methods that draw directly on pairwise comparison, e.g. (Gonnet et al. 1992, Watanabe & Otsuka 1995, Koonin et al, 1996, Harris et al. 1992, Tatusov et al. 1997, Krause & Vingron 1998). All these works induce a clustering of the input database, based on the transitive closure of similarity scores (i.e. single linkage clustering). Among these works, two (Harris et al. 1992, Tatusov et al. 1997) have addressed the problem of multi-domain proteins. Harris et al. allow groups to merge only if they share k overlapping regions. However, they concluded that k=1 is the best choice for maximum accuracy. Thus their clustering procedure essentially remained a single-linkage clustering (in multi-trait proteins, regions are classified to multiple classes). In the second study (Tatusov et al. 1997), clusters are created starting from triangles of homologous proteins from different lineages. Triangles which share an edge are merged (this requirement reduces the probability that unrelated clusters merge). An additional (manual) step is carried to split clusters which are incorrectly merged due to multi-domain proteins.

Conceptual difference: The important role of motifs and domains in defining protein's functionality is unquestionable. Detecting a known motif within a new protein sequence can help reveal its function and lead to the correct assignment of the new sequence to an existing protein family. Indeed, domain-based studies have added

much to our knowledge. However, in many cases, characterizing a new protein only by its domain content, seems insufficient. This happens, for example, when no known domains are apparent in the new protein. In some instances, only few related sequences are available, too few to define a reliable prototype signature or a profile of the common domains. Therefore, a proper analysis of a new protein sequence should incorporate comparisons against domain based databases, as well as sequence databases. In this view, an analysis which identifies groups of related proteins in databases of protein sequences is valuable. It may amplify the outcomes of an exhaustive pairwise comparison. When close hits are already grouped together based on their mutual similarity, this may highlight a similarity with a group which could otherwise be missed by a simple manual scanning. Moreover, if several groups are found related to the query sequence, this may indicate the existence of several distinct functional/structural domains. If all groups share the same region of similarity, this adds insights about the relations between the different groups. In some cases, this may suggest that the groups belong to the same family or superfamily. Some of this information, but not all of it, can also be inferred from comparisons with domain databases. The domain-based databases usually offer a lot of detailed information about domains and the domain structure of proteins, through multiple alignments, and schematic representations of proteins. This information is of great value for biologists. Moreover, unlike domain based methods, protein-based methods can be applied to all protein sequences and are more easily automated.

The present invention belongs to the second category. It draws on pairwise similarities and looks for strongly connected sets of proteins. It applies a moderate version of transitive closure, in an attempt to eliminate chance similarities and avoid indirect multiple-domain-based connections.

When begining with a very strict, high-resolution classification that employs only connections of very high statistical significance, clusters are merged to form bigger and more diverse clusters. The algorithm operates hierarchically, each step considering weaker connections in addition to the connections that are already accounted for. A statistical test is applied in order to identify and eliminate suspicious/problematic connections as well as possibly false connections between

unrelated proteins. The output is thus an hierarchical organization of all protein sequences. The algorithm incorporates no further biological information and uses only the similarity scores that are provided by standard methods.

The method described here applies to the set of all SWISSPROT (Bairoch & Boeckman 1992) sequences, and yields an exhaustive classification. This approach leads to the definition of a new pseudo-metric on the space of all protein sequences. In most cases, this emerging metric turns to be more sensitive than the existing measures on the basis of which it was derived. Such metrics are necessary in the quest of a global self organization of all protein sequences, as discussed in (Linial et al. 1997).

Methods

This section contains a description of the computational procedure of the present invention. The procedure was carried out on the SWISSPROT database (Bairoch & Boeckman 1992) release 33, with a total of 52205 proteins.

Defining the graph

The space of all protein sequences is represented as a directed graph, whose vertices are the protein sequences. Edges between the vertices are weighted with weights that reflect the dissimilarity between the corresponding sequences, i.e. high similarity translates to a small weight. To compute the weight of the directed edge from A to B, one compares A against all sequences in the SWISSPROT database, and obtains a distribution of scores. The weight is taken as the expectation value (Altschul et al. 1994) of the similarity score between A and B, based on this distribution. This is an estimate for the number of occurrences that the appropriate score could have been obtained by chance, i.e. when compared with random sequences drawn from the same background distribution (usually defined as the distribution of amino acids overall the database). A low expectation value reflects a strong connection, of high statistical significance, and a high expectation value entails an insignificant, weak

connection (expectation value decreases as the length of the similar region decreases, and self similarity of short proteins is less significant then self similarity of long proteins. Therefore, when distance function between protein sequences is sought, the expectation values may need to be corrected for the query length). Not all edges are retained in the graph. Edges of statistically insignificant similarity scores get discarded (details below). In other words, in the final graph, an edge between sequences A and B indicates that the corresponding proteins are likely to be related.

This graph has been constructed, using the common algorithms for protein sequence comparison; Smith Waterman dynamic programming method (SW) (Smith & Waterman 1981), FASTA (Lipman & Pearson 1985) and BLAST (Altschul et al. 1990). The SW algorithm was run with the BLOSUM62 matrix (Henikoff & Henikoff 1992) and gap penalties of -12,-2 using the Bioccelerator hardware (Compugen) or the 'ssearch' program which is part of the FASTA 2 package by Bill Pearson. FASTA was run using the 'fasta' program with the BLOSUM50 matrix (Henikoff & Henikoff 1992) and gap penalties -14,-2 (the default setting). Both 'ssearch' and 'fasta' calculate expectation values based on empirically derived distribution of scores (Pearson 1998) (the Bioccelerator apply the same procedure for assessing the significance of results, as in ssearch). The BLAST algorithm was run with the BLOSUM62 matrix using the blastp 3 program available from the NCBI ftp site. The program reports similarity scores along with the probability that the scores could have occurred by chance. All these methods are in daily use by biologists, for comparing sequences against the databases. Though SW tends to give the best results on average, it is not uncommon that FASTA or BLAST are more informative, especially when combined with different scoring matrices (Pearson 1995). Therefore all three methods were incorporated into the graph, to achieve maximum sensitivity.

The following sections contain a detailed description of the procedure of assigning weights to edges. The procedure starts by creating a list of neighbors for each sequence, based on all the three methods. In order to place all three methods on comparable numerical scales, a numerical normalization is applied first to all methods. Then, only statistically significant similarities are maintained in these lists. Finally, the

weight of an edge is defined as the minimum associated to it by any of the three methods, to capture the apparently strongest relation.

Placing all methods on a common numerical scale

It is relatively easy to compare between scores that a particular method assigns to different comparisons. However, how does one compare between scores that are assigned by different methods? The following calculation was performed: Pick any protein, carry out an exhaustive comparison against the whole database and consider the highest scores in each of the methods. Now plot these values and compare two methods at a time. These Scores show a remarkably strong linear relation in log-log scale (Fig. 16), therefore by introducing a (usually small) correction factor, per each protein and per method, the three methods get scaled to a single reference line.

Fig. 16 is a graphic representation of the correlation of BLAST e-values and SW e-values. (a) BLAST e-values of neighboring sequences of 1431_lyces (P42651) vs. the SW e-values of the same neighbors. The graph is plotted in log-log scale. Note the strong linear correlation between the scores assigned by the two methods, where the correlation coefficient is 1.16, i.e. e-value BLAST= $(evaluesw)^{1.10}$. The fitness of the linear line is even better when the line is of the form a * x + b. For 1431_lyces the best fit achieved by 1.1*x-5.5, i.e, evalueBLAST =: $10^{-5.5} * (evaluesw)^{1.1}$ (b) BLAST e-values of neighboring sequences of la03-human (PO4439) vs. the SW e-values of the same neighbors. Note that the correlation coefficient is 1.3 in this case (the improvement of the fitness by introducing the 'b' term in the linear fit is marginal in this case).

Defining the list of neighbors

It is, of course, very difficult to set a clear dividing line between true homology and chance similarity. Expectation values below 10^{-5} can be safely considered significant and those above 10 reflect almost pure chance similarities. However, the midrange is difficult to characterize, and truly related proteins may have expectation values around 1. An overly strict threshold will miss important similarities within the twilight zone,

whereas an excessively liberal criterion will create many false connections. The exact threshold for each pairwise comparison method was set to best discriminate among related and unrelated proteins. The choice is based on the overall distribution of expectation values over the entire protein space, as given by each of the three methods.

This is illustrated in Figs. 1A-1C, which shows the distribution of expectation values over the entire SWISSPROT database, for SW, FASTA, and BLAST, based on the neighbors lists of all protein sequences in the SWISSPROT database. The distribution may be thought as the average distribution of expectation values, for a 'typical' protein sequence as a query. The graphs in Figs. 1A -1C naturally suggest a threshold for each method. The distribution drawn in a log-log scale is nearly linear at low expectation values, but starts a rapid increase at a certain value. The steep slope indicates a rapid growth in the number of sequences that are unrelated to the query sequence (high expectation values). The distribution may differ from one sequence to another, and accordingly the threshold may change. However, the number of neighboring sequences of a single protein is usually not enough to deduce ā reliable threshold. Only when the distributions are averaged then a general threshold is obtained, which can be used a s a guideline.

In this view, the threshold was set at the value where the slope rapidly changes. The thresholds for SW, FASTA and BLAST are set at 0.1, 0.1 and 10^{-3} respectively. An edge from vertex A to vertex B is maintained only if a significant score is obtained on comparing the corresponding proteins. Namely, if either SW or FASTA yield an expectation value ≤ 0.1 or BLAST's expectation value is $< 10^{-3}$.

While the self-normalized statistical estimates of FASTA and SW (Pearson 1998) are quite reliable (see also Brenner et al. 1998), the statistical estimates of BLAST may be effected by the amino acid composition of the query sequence, and an unusual composition (e.g. low complexity segments within the sequence) may strongly bias the results of a search (Altschul et al. 1994). Therefore, the results of BLAST were consulted following a filtering of the query sequence, to exclude low complexity segments, using the SEG program (Wootton & Federhen 1993). Consequently,

filtering may significantly reduce the number of high scoring hits reported by BLAST. However, if only sequences that pass the filter are acknowledged, many relations of biological significance may be missed. Instead, a more stringent threshold is set in this case for BLAST at 10^{-6}.

A major difference between BLAST and SW/FASTA is that BLAST does not count/penalize for gaps (The new version of BLAST does account for gaps in the alignment. This heuristic is a very good approximation to the SW algorithm, whose main advantage is in its speed. It is, however, inferior to the rigorous SW algorithm. In contrast, the old version of BLAST occasionally detects similarities that are missed by SW (e.g. for the Glucagon precursor family, and the H'-transporting ATP synthase [Pearson, 1995]).. BLAST defines the similarity based on one or more high-scoring segment, pairs (ungapped local alignments), and the significance is assessed by applying Poisson or sum statistics (Altschul et al. 1994). Consequently, since gaps are ignored, BLAST tends to overestimate the statistical significance of fragmented alignments, and this behavior of BLAST is countered by the above asymmetry in selecting the edges. While this property may help BLAST reveal significant similarities that the other methods miss (e.g. Pearson 1995), it is cautious to be aware of highly fragmentary alignments that cannot be considered biologically meaningful. Therefore, those BLAST scores that come from a large number of HSPs (high scoring pairs), when the MSP (maximal segment pair) is insignificant are ignored.

Finally, even if the comparisons between proteins A and B fail to satisfy the previous criteria, the edge from A to *B is* maintained when all three methods yield an expectation value ≤ 1.

This procedure is designed/helps to screen most of the chance similarities in the neighbors list of each protein sequence. Unfortunately, chance similarities may occasionally pass the set criteria. A major goal of the algorithm that is described next is to detect such similarities and eliminate them.

Exploring the connectivity

In exploring the graph clusters of related sequences which hopefully have a characteristic biological function are sought.

There are two major obstacles which should be considered: i) Multi domain proteins can create undesired connection among unrelated groups; ii) Overestimates of the statistical significance of similarity scores may bias decisions. Naturally, chance similarities become more abundant as significance levels decrease.

Therefore, transitivity should be applied restrictively. By analogy, if transitivity is to be viewed as a force that attracts sequences, then it should be countered by some "rejecting forces" so that unrelated clusters be kept apart and prevent a collapse in the protein space (Though at the level of 10^{-100} this does not occur).

The approach

The approach is to begin by eliminating all edges of weight below a certain, very high, significance threshold (i.e. low expectation value). This operation splits our graph to many small components of strong connectivity. In biological terms, we split the set of all proteins into numerous small groups of closely related proteins, which correspond to highly conserved sub families.

To proceed from this basic highly restrictive classification, the threshold is raised, in a stepwise manner, and more relaxed statistically significant similarities are taken into account. In so doing, several clusters of a given threshold may merge at a more permissive threshold. However, this process is closely monitored and a merge is allowed only when strong statistical evidence is found for a true connection among the proteins in the resulting set.

Basic classification

If all edges of weight below a certain significance threshold are eliminated, the transitive closure of the similarity relation among proteins splits the space of all protein sequences into connected components or clusters. These are proper subsets of the whole database wherein every two members are either directly or transitively related. These sets are maximal in this respect and cannot be expanded. Thus they offer a

self-organized classification of all protein sequences in the database. The threshold is set at the very stringent significance level of 10^{-100}. Similarities which are reported as significant above the level of 10^{-100} are very conserved and stretch along at least 150 amino acids. Thus, neither chance similarities, nor connections based on a chain of distinct common domains in multi-domain proteins occur at this level . The resulting connected components can be safely expected to correlate with known highly conserved biological subfamilies.

Note that this is a directed graph, and hence is not necessarily symmetric. Specifically, it may happen that there is an edge from protein A to protein B, but none in the reverse direction (a directed graph is strongly connected if for every two vertices there is a directed path from x to y as well as from y to x.) Furthermore, even if both edges exist, their weights usually differ. Therefore, the notion of a component is that of a strongly connected component. The partition into strongly connected components is clearly more refined that the partition into connected components.

The clustering algorithm

The procedure is recursive. That is, given the classification at threshold T, a method for deriving the classification at the next more permissive level should be given, that is 10^5 - T. The start is from the basic classification at 10^{-100}.

Figure 14 describes the clustering algorithm. Phase 1) Identify pairs of clusters that are considered as candidates for merging. Decisions are made based on the geometric mean of the pairwise scores of the connections between the two clusters. If this mean exceeds a specific threshold then the cluster is accepted as a candidate, and enters a pool of candidates. Otherwise it is rejected (denoted in the figure by 'X'). Phase II) pairwise clustering is applied to identify groups of clusters which are strongly connected. At each step the two closest groups are chosen from the pool and merged provided that the geometrical mean of all pairwise similarities pass the threshold. Otherwise they stay apart (denoted by dashed line).

The algorithm runs in two phases. First groups ("pools") of clusters that are considered as candidates for merging are identified and marked (see Fig. 14). A local

test is performed where each candidate cluster is tested with respect to the cluster which "dragged" it to the pool, to check their mutual similarity.

To quantify the similarity of two clusters P and Q, the geometric mean of all pairwise scores of protein pairs one of which from P and the other from Q, are calculated. Unrelated pairs are assigned the default (insignificant) e-value of 1. The geometric mean reflects the distribution of pairwise connections between the two clusters, such that random or unusual connections have only a little effect. When the geometric mean of the e-values is below $\sqrt{T}$ (more significant) the interpretation is that P and Q are indeed related and that their connection does not reflect chance similarities or chain of domain-based connections. The level of confidence in the reliability of the connection clearly decreases as T increases. The Quality of the P - Q connection is defined as minus the log of the geometric mean. This quantity ranges between 0 and 100, and the higher it is, the more significant the connection.

At the end of the first phase a group of clusters is left which are candidate for merging. This group is the input for the second phase.

At the second phase a variant of a pairwise clustering algorithm is carried out. This algorithm successively merges only pairs of clusters that pass the above test and are not suspected as representing chance or domain-based similarities. At each step the two closest clusters are chosen (based on the quality of their connection) and merged if their similarity (as quantified above, and based on all pairwise similarities between the new formed clusters) is more significant than the threshold. The process stops and the final clusters are defined (Fig. 14) when the similarity of the next closest clusters do not pass the threshold.

All the rejected merges are marked and registered for further biological analysisTthese rejected merges are refered to as possibly related clusters.

This analysis is performed at different thresholds, or confidence levels, to obtain an hierarchical organization. The analysis starts at the 10^{-100} threshold. Subsequent runs are carried out at levels 10^{-95} , 10^{-90} , 10^{-85} , 10^{-80}. The process terminates at the threshold of 10^{-0} =1. Above the threshold of 1 almost all similarities are in fact chance similarities (see previous section).

Results

Nearly all of the clusters found are biologically meaningful, some corresponding to well known families, and many others represent less studied families. Some clusters exclusively consist of unknown proteins or hypothetical proteins.

The essence of this work is a fully automatic classification of all protein sequences, without incorporation of any biological considerations other than the use of standard pairwise sequence comparisons. The analysis concerns complete proteins, and is not limited only to those subsequences which are identified as functionally or structurally important motifs and domains. Indeed, not all the emerging clusters are correlated with a specific domain, although some of the clusters encountered are characterized by a domain that is common to many or all member proteins.

General information

Figure 19 shows a table showing the distribution of cluster sizes at various confidence levels. At each level, the set of all proteins splits into clusters, which merge to form larger and coarser clusters as the confidence level decreases. In particular, the number of isolated proteins (clusters of size 1) diminishes as well. At the lowest significance level there are 10.602 clusters, of which 4435 contain at least 2 members. 1006 clusters have size 10 and above.

The number of clusters (of size bigger than 1) at each level of confidence ranges between 4228 and 5543. How many clusters should there be ? A lower bound for this number is provided by the number of different folds, since it is generally expected of members of the same cluster to have similar folds. The current estimates place the number of folds between several hundreds and one thousand (Wang 1996, Chothia 1992). However, the same fold is usually adopted by few different superfamilies which share little or no sequence similarity. Therefore, in a sequence based analysis it should probably be expected that these superfamilies correspond to different clusters, possibly related/connected. Moreover, superfamilies may consist of several families, sometimes with as little as few percentages of sequence identity. Consequently, these families may be classified to different clusters, depending on the method sensitivity. All

together the total number of clusters is expected to exceed the number of folds, but it is quite safe to predict that by no more than a whole order of magnitude. A number of clusters in the 4000-6000 range is, thus, consistent with this estimation.

Almost each of these thousands of clusters is biologically meaningful. Many of the smaller clusters suggest a definition of new biological families, and often a characteristic feature is evident. Figure 15 shows only the 50 largest clusters at the lowest confidence level (10^{-0}) that is considered. The description attached in the table of Fig. 15 to each cluster is based mainly on the SWISSPROT annotations of its members. The table should be viewed only as a sample. Henceforth, unless otherwise stated, cluster numbers refer to level 10^{-0}.

As is apparent from Fig. 15, largest clusters are at the lowest confidence level (10^{-0}). (i) Clusters are ordered in decreasing order of size (ii) The order of transitivity within each cluster is defined as follows: take one of the proteins in the cluster as the cluster's seed, The seed's order of transitivity is 0. Its neighbors are of order 1. Additional proteins that are neighbors of 1st order proteins, are of order 2, etc. The definition does not depend on the choice of the seed protein. (iii) The family description states the main feature of the member proteins.

Performance evaluation

It is very hard to evaluate the validity of classifications that merge from a large scale study of protein sequences. No generally accepted standards have been set yet in this field. Thus over the years, new classifications were traditionally compared with what is considered (one of) the state of the art characterization of protein sequences, namely, the PROSITE dictionary of signature patterns, motifs, and domains (Bairoch et al. 1997). For domain based studies, a comparison with the manually derived PROSITE dictionary is inevitable, and essential in order to verify that the results are biologically meaningful. However, when the analysis is not limited to regions which are known or suspected as domains, no standard benchmark exists to assess the

quality of the results. One may then resort to domain-based databases for a guide-line. Obviously, this may bias the assessment, and should be kept in mind when evaluating the results.

To estimate the quality of the classification it is compared with two databases, PROSITE and Pfam (both domain-based databases), using four indices of quality.

The evaluation methodology

Given a reference classification A and a new classification B of the same set *X*, the quality of the classification U is evaluated in terms of the reference classification A, by means of their mutual agreement . The two classification can be either "hard", i.e. each protein is classified to exactly one group, or "soft" where each protein can be classified to more then one group. In the present case, Pfam and PROSITE are soft, while the present is a hard classification .

Gracy and Argos (Gracy & Argos 1998) have proposed a procedure for performance evaluation. Each class a *E A* is associated with the group b *E* B which maximizes the quantity *tp-fp-f n*. Here true positives (tp) are given by /anb/ - 1, false positives (fp) are given by ib \ al and false negatives (fn) are given by la \ bl. Quality is defined by the percentage of the true positives 100 – tp/tp+fp+fn.

In the same way, the percentage of false positives and false negatives are calculated. To simplify the comparison between the present results and their results, this index is used, denoted Qsingle.

The procedure resembles the one suggested by Gracy and Argos. The difference is that each group a *E A* is associated with one or more groups b from B, where each group b satisfies tp > *fp* . Specifically, that a group b *E B* is a relative of a group a E *A* if more than 50% of its members are also members of a (see figure Fig. 13). For each group a E *A* all its relatives b in *B* are identified. The union of all the relatives of a is denoted by b_a. A protein is misclassified by classification B if it is a member of a missed by b_a (false negative), or is a member of b_a ,but not a member of a itself (false positive). The conjunction of b_a and a defines the group of proteins which are correctly classified.

Figure 13 schematically shows association of groups in classification B with groups in the reference classification A. Groups B I- B 4 are relatives of the ellipse (group of A), while groups B5 and B6 are not.

The quality Qset of the classification for the group a is defined by the percentages of the true positives (proteins which were correctly classified) out of the total number of proteins in a and b_a, i.e. $100.\frac{a \cap b_a}{a \cup b_a}$

This measure accounts for both error types. This procedure is repeated for every group a E A, and the total percentage of true positives is given by the average overall groups a E A.

Since dividing a family a into many small clusters (and in the extreme, to singletons clusters) is not desirable, another quality index is also defined, that accounts for the number of clusters which are relatives of a, and decreases as this number increases. This is done by first subtracting the number of relatives of a from the number of true positives, then calculating the percentages out of the total number of proteins in a and b_a, i.e. $100.\frac{|a \cap b_a| - \text{number of relatives of } a + 1}{a \cup b_a}$

This way a family of size *N* which is mapped to a single cluster of size *N* has quality of 1 (or 100%), while a family of size *N* which correspond to N singletons in the present classification has quality 0. This index is denoted by Qset-relatives.

The fourth index of quality is given by the quantity *tp/(tp+fn)*, which is the portion of the reference family a which is in the set Rel_a. This measure does not take into account the false positives. The reason it is used is that not all false positives are indeed false positives (as shown in the next sections). This way all false positives are assumed to be potential related sequences. Obviously, this is not true. However, it is useful as an upper bound on the quality of the classification, assuming that at least some false positives are essentially potential related sequences. It is denoted by Qupper-bound

A consistency test was proposed by Krause and Vingorn (Krause & Vingorn 1998). Although their test can be used for rough self-validation, it is less useful for assessing the quality of a new classification, with respect to a reference classification.

The reference databases

The present classification was compared with two domain based classifications: PROSITE and Pfam. The present classification contains 52,205 proteins of the SWISSPROT database, release 33, classified to 10,602 clusters (at the lowest level of the analysis), of which 1006 are of size 10 and above (see Fig 19). The PROSITE database release 13 (released/associated with the SWISSPROT 33 database) contains 24156 proteins, characterized by 1151 different signature patterns. It is very common in PROSITE that the same family is characterized by few signature patterns. Therefore, all patterns that are documented the same in PROSITE are considered to be the same family. The exceptions are those who never appear together in the same protein though documented the same. Also considered different (for the purpose/in the scope of assessment) are those groups that differ by more than 10 proteins. Also, patterns which are always associated with other patterns (i.e. the corresponding groups are completely included in bigger groups) are ignored though documented differently in PROSITE (e.g. INTEGRIN RETA in EGF, POU-1 and POU-2 in HOMEOBOX). Overall, within these terms, the 1151 signatures characterize 874 protein families and domains, of which 600 are of size 10 and above. The Pfam database, release 1.0 (associated with SWISSPROT. 33) contains 15604 proteins, classified to 175 families, of which 172 are of size 10 and above.

Evaluation results

The results of the evaluation procedure for these reference databases are given in table in Figure 20. The evaluation is based on all families with at least 10 members (the same analysis with all families with at least 5 members gave the same results up to deviation of less then 1%).

In (Gracy & Argos 1998) a similar procedure with respect to a database which is a combination of PROSITE and PIR (George et al. 1996) resulted in 96.6 % true positives (1.8% false positives) in PROSITE (which means that the combined database is very close to the PROSITE database), 93.2 % true positives in DOMO

(0.3% false positives) and 65.1% true positives in ProDom (0.9% false positives). All three are domain-based classifications. The first three numbers in the table of Fig. 20 assess the performance of the method of the present invention in the same way as the procedure in (Gracy & Argos 1998) and therefore are approximately in the same scale. With 69% true positives, the present work compares favorably with ProDom, though false positives are more abundant. However, not all of these false positives should be counted as false positives (see next section).

Since the present is a hard classification, the second measure Qset is found to be more appropriate. This means that no single cluster is associated with one domain family. The different clusters may correspond to different subfamilies, but the overall structure is still detectable through connections between clusters (see section 'possibly related clusters'). An example is the PROSITE family AA_TRNA_TIGASE.I family. Seven different clusters form the cover of this family corresponding to the subfamilies: leucyl/isoleucyl/valyl/methionyl-trna synthetase (cluster 152), glutamyl/prolyltrna synthetase (cluster 429), tryptophanyl-trna synthetase (cluster 758), arginyl-trna synthetase (cluster 988), tyrosyl-trna synthetase (cluster 904), cysteinyl-trna synthetase (cluster 1216), and a singleton (cluster 6647) arginyl-trna synthetase (a very short fragment). With this measure the quality of performance reaches 76.7%.

Drawbacks of the evaluation methodology

The above procedure may result in an over-strict measure, and a few factors that may suppress the performance need some careful inspection. One such factor is the amount of false positives. In many cases the supposedly false positives are actually true positives. For example, short fragments which are undoubtfully identified as belonging to a specific family may be considered false positive simply since they are too short and do not share the domain which is common to all other members in the family. Similarly, hypothetical proteins which share significant sequence similarity with a family are not necessarily false positives. Even proteins which are documented as belonging to a family, may be counted as false positives simply since they do not have exactly the family signature pattern, but rather a slightly modified one. For example,

cluster 27 has 139 proteins out of which 132 has the actins and actins like signature. Five of the other 7 proteins are documented as actins (and indeed show a remarkable similarity with other actins) and two are hypothetical proteins (again, with a strong similarity to actins). However, these 7 proteins do not have the actins and actins-like signature and therefore are counted, unjustly perhaps, as false positives. Similarly, cluster 7 has 46 proteins which do not have the rubisco large signature, and counted as false positives, but all are annotated in SWISSPROT as ribulose bisphosphate caxboxylase large chain (some are fragments). In the same way cluster 15 has 13 proteins which do not have the cytochrome p450 signature, 10 of which are variants of cytochrome p46O, 2 proteins are thromboxane-a synthase, and one is trans,-cinnaniate 4-monooxygenase., all are documented to be part of the cytochrome p45O family, and indeed, show a strong similarity with cytochromes p460.

False positives of these types are very common in the present clusters. Should they be counted as false positives ? Obviously, some hits are real false positives, but there is no automatic way to detect which false positives are biologically meaningless and which are not. It can only be assumed that some false positives can be ignored. Consequently, the performance is expected to improve. An upper bound of the performance is given by the third index of quality (last column). With Qupper-bound = 87.4% true positives. The actual quality is expected to be somewhere in between the 76.7% given by Q,set and the 87.4%.

Another factor, which may drastically decrease performance, is that some families were only characterized for subfamilies while other members in the family, are not known to share a well defined domain. In the evaluation procedure they will be considered as false positives, and since they may dominate then no cluster will be matched with the subfamily. One such example is the *ran* family which is a subfamily of the small G-proteins. The *ran* proteins are classified to the same cluster as the ras/ras-like/rab proteins (cluster 12). However, since the ras/ras-like/rab proteins are not characterized by the same signature pattern (nor are they characterized by any other signature pattern in release 13.0 of PROSITE), then they are considered as false positives. While this may be correct for the *ran* subfamily, by definition, the real

functional relationship with all the other proteins in this cluster may suggest that the definition of false positive is somewhat too strict.

The families which are difficult to resolve in the present analysis (families for which performance is worst, e.g. less then 50% true positives) are dominated by short/local domains (e.g. PH domain, EGF, ER-TARGET, C1Q, KRINGLE, C2 domain, SH2, SH3) or domains that are paired with other, more abundant domains (e.g. opsin paired with G-protein-receptor). This is expected, since the present analysis is not a domain-based.

The results of the evaluation procedure with Pfam as the reference database, lead to approximately the same conclusions: good performance for protein families, but short motifs are not detected well. The quality of the classification increased as the coverage (the total portion of the sequences which was included in the multiple alignment used to define the domain or the family in the Pfam database) increased, and for coverage > 0.3 (134 families) the quality raised to Qsingle = 76.9% (6.1% false positives), Qset = 84.9% (6.2% false positives) and Qupper-bound = 90.9%. For coverage > 0.5 (109 families) the quality raised to Qsingle, = 80.6% (5% false positives), Qset = 88.3% (5% false positives) and Qupper-bound = 93.3%.

New clusters

The above evaluation procedure does not take into account the many new clusters in the present classification which do not have a counterpart cluster in PROSITE, nor in Pfam. Out of the 1006 clusters with more than 10 members each (total of 33682 proteins), 308 clusters (6989 proteins, which are 20.8%) do not have a counterpart family in PROSITE database. 734 clusters (15586 proteins 46.3%) do not have a counterpart in Pfam A. 281 clusters are missing from both.

The largest 20 unannotated clusters (both by PROSITE and Pfam) are listed in the table of Fig. 21. They are documented based on their SWISSPROT definition. The purity of these clusters (by means of definition consensus) is quite high. Still proteins in these clusters are not characterized in PROSITE nor in Pfam.

The method of the present invention as an analysis tool

Tracing the formation of clusters

A major aspect of the hierarchical organization is that clusters at a given threshold may merge at a more permissive threshold. This reflects the existence of subfamilies within a family, or families within a superfamily.

By moving from one level to a more restrictive one, a subdivision of clusters into smaller subsets is obtained. These subsets suggest a natural division of the corresponding family, as illustrated in the following example for the transport system permease proteins.

Figure 12 is a schematic illustration tracing the formation of clusters. Cluster 170 - the transport system permease proteins. In moving on to level 10^{-5} and further to level 10^{-10} the cluster splits into few subclusters. Each circle stands for a cluster at threshold 10^{-10}. Circles' radii are proportionate to the clusters sizes (numbers indicate clusters' sizes). The drawn edges appeared upon changing the threshold from 10^{-10} to the more permissive 10^{-6}. Edge widths are proportionate to the number of connections between the corresponding clusters.

Cluster 170 at level 10^{-0} with 46 proteins consists of transport system permease proteins. These proteins participate in multicomponent transport systems in bacteria. Specifically, they are the integral inner-membrane proteins which translocate the substrate across the membrane.

The cluster decomposes into four subclusters at level 10^{-10}, which form a clique as in Fig. 12. These smaller subclusters correspond to the lactose/maltose transport system lacG/malG, the lactose/maltose transport system lacF/malF. the phosphate transport system, and other transport systems (of sulfate, molybdenum, spermidine and putrescine). The subgroups of lacG/malG and lacF/malF form already at level 10^{-25}. Some proteins that combine features from F and G subtypes are denoted in SWISSPROT as malGF proteins. However, based on this subclassification and the fact that the malG group and the malF group form at such high levels of significance, these proteins may be classified either to malG or to malF.

Hierarchical organization within protein families and superfamilies

The hierarchical organization also suggests classification within known, families. This classification is suggested by scanning the hierarchy over all levels, as illustrated for the small G-protein/Ras superfamily.

The ras gene is a member of a family that have been found in tumor virus genomes and are responsible for the viruses' carcinogenic effect. In most cases this viral oncogene is closely related to a cellular counterpart, called proto-oncogene. infection by a retrovirus that carries a mutant form of the ras gene (ras oncogene), or mutation, can cause cell transformation. Indeed, mutations in ras gene are linked to many human cancers.

The cellular ras protein binds guanine nucleotide and exhibits a GTPase activity. It participates in the regulation of cellular metabolism, survival and differentiation. In the last decade many additional proteins that are related to ras were discovered, all of which share the guanine nucleotide binding site. They are referred to as the small-G-protein superfamily (Nuoffer & Balch 1994). This family of proteins has several subfamilies: ras, rab, ran, rho, ral, and smaller subfamilies. Like ras, these proteins participate in regulatory processes,such as vesicle trafficking (rab),and cytoskeleton organization (rho).

Figure 17 is a schematic representation of the hierarchial organization of the small G-protein family. This family is composed of several subfamilies. A total of 229 proteins, combined in cluster 12 (level 10^{-0}), were grouped together into isolated sets at different levels of confidence, to form a natural subclassification of the family. This hierarchical organization is much enriched by combining possibly related clusters. The related clusters are connected by dashed lines.

In Fig. 17 the relations within this family are depicted based on the present hierarchical organization. A total of 229 proteins, all from the small G-protein superfamily, are presented. All were clustered into cluster 12 at the lowest level of significance 10^{-0}. Small clusters, which correspond to subfamilies, are formed at higher confidence levels, and fuse to larger clusters when the threshold is raised. The four main branches coincide with (I) rab subtypes (II) ras, ral and rap (III) rac, rho and cdc (cell division control proteins) (IV) ran. Interestingly, the linkage of rac to cdc

and rho seems stronger than that between ran and rho or rab and rho. This proposed subdivision suggests a common root for all the subtypes, but splits them in a way that resembles the evolutionary tree of the small G-protein superfamily (Downward 1990).

As include weaker similarities are included, other families are identified which are related to the small G-protein family. According to our map, the clusters which are detected as related clusters include cluster 29 (138 proteins), cluster 646 (14 proteins) cluster 461 (20 proteins) and cluster 1400 (7 proteins). All these clusters are possibly -related clusters (rejected merges) of cluster 12 (see next section for a discussion on possibly related clusters). Cluster 29 consists of ADPribosylation factors family (APF) that are involved in vesicle budding, and of guanine nucleotide binding proteins from the sar subfamfly, whose members participate in a different type of vesicle budding. Most interestingly - Gα proteins of heterotrimeric G proteins belong to the same cluster.

The connection between this cluster and cluster 12 is based on similarity of the ARFs and the rab subfamily, as shown in Fig. 17. Cluster 646 consists of the GTP-binding protein ERA and of thiophene/furan oxidation proteins (both being groups of GTP-binding proteins). This cluster and cluster 12 are related through the similarity of the thiophene/furan oxidation proteins and the ras subfamily. Cluster 461 (GTP-binding proteins of the OBG family) and cluster 1400 (hypothetical small G-proteins) are not directly related to cluster 12. However, these clusters are related to clusters 29 and 646, as well as to each other (see Fig. 17).

Possibly related clusters and local maps

The clustering algorithm automatically rejects many possible connections among clusters. This happens whenever the quality associated with this connection falls below a certain threshold (see Fig. 14). Many of these rejected connections are nevertheless meaningful and reflect genuine though distant homologies. The rejected mergers are referred to as possibly related clusters.

In examining a given cluster, much insight can be gained by observing those clusters which are possibly related to it. Even though some of these connections are justifiably rejected, in particular at the lowest level of confidence (10^{-0}), many others do reflect structural/functional similarities, despite a weak sequence similarity. Based on the connections with possibly related clusters we can plot local maps (at this stage, mostly schematic) for the neighborhoods of protein families. These schematic maps can expose interesting relationships between protein families. Here we present a map for the immunoglobulins superfamily.

Two of the big clusters in the tables of Fig. 11 belong to the immunoglobulin superfamily. These are cluster 4 (immunoglobulin V region) and cluster 5 (immunoglobulins and major histocompatibility complex).

Fig 11A shows clusters possibly related to cluster 4 (Level: 1e-0). Clusters are sorted by quality (i.e. minus the log of the geometric mean of similarity scores). Note that all clusters belong to the superfamily of immunoglobulins.

Fig. 11A shows those clusters which are possibly related to cluster 4, ordered by their quality value. These clusters include proteins which are involved in aspects of recognition at the immune system via the variable regions.

Fig. 11B shows clusters possibly related to cluster 5 (Level: 1e-0). Only clusters with two members or more are shown, and are sorted by quality. Almost all clusters belong to the superfamily of immunoglobulins.

Clusters 596 and 866 are probably unrelated (in italic).

Likewise, the table of Fig. 11B shows the clusters which are possibly related to cluster 5. These clusters, unlike the clusters related to cluster 4, consist mostly of proteins that adopted the immunoglobulin fold of the Ig constant region. Clusters which are suspected to be unrelated appear in italic (one can validate the significance of "possibly related clusters" using the quality of their relation and the alignments, and insignificant connections can be easily traced and ignored by a manual examination of the alignments).

The two sets of clusters are mostly disjoint, with the exception of cluster 1796. Members of this cluster contain both regions, whence the cluster is related to both

clusters 4 and 5. The different parts of the proteins account for the appropriate relationships.

Fig. 18 is a schematic map of the immunoglobulin superfamily. All clusters related to clusters 4 and 5 are shown (these are referred to as directly related clusters), as well as other clusters which are indirectly related clusters (clusters that are possibly related to directly related clusters). In this schematic map each cluster is represented as a circle, whose radius is proportional to the cluster's size. Only clusters' numbers are given. See Fig. 22 for more information on these clusters. Clusters on the left contain an Ig V-region. The group on the right has an Ig C-region. The clusters in between share both regions (except for cluster 1468, that consists exclusively of V-region).

There is also a direct connection between cluster 4 and cluster 5. The connection is based on 226 pairwise similarities between cluster 4 and 5. However, all the similarities are due to a single protein in cluster 5, a T-cell receptor beta chain (PII364). This protein has one V region aside of a C region, and the similarity with the 226 proteins in cluster 4 is limited to the V region. No other protein in cluster 5 has a V region. Note that despite these similarities. clusters 4 and 6 did not merge, and the connection was automatically rejected.

Figure 22 is a table showing clusters belonging to the immunoglobulin superfamily. All clusters appear in the local map of the immunoglobulins (Fig. 18).

This information depicts the 'geometry' of the protein space in the vicinity of the immunoglobulin superfamily as in Fig. 18. This map includes also indirectly related clusters, i.e. possibly related clusters of order 2 and above (related clusters of related clusters, etc). The map includes almost all protein families which belong or are related to the immunoglobulin superfamily (see Fig. 22 for details), except for three clusters: cluster 373 (isolated cluster of periplasmic pilus chaperones), cluster 2172 (isolated cluster of THY-1 membrane glycoproteins) and cluster 2363 (B-lymphocyte antigen cd19).

This "local map" of relationships is plotted to distinguish between three main groups. The left one corresponds to the variable regions (V domains) of the immunoglobulins, the right one to the constant regions (C domains). In the middle appear many clusters that are a mixture of the two types.

The evolutionary pathway between the structural classes of Ig and Ig-like domains are not yet established (Smith & Xue 1997). However, the many alternative connections between clusters whose proteins resemble V-domain to these of the C-domains indicate that adhesion molecules, fasciclin II and vascular CAM, are positioned between the classical V and C-regions. Indeed, these non-Ig molecules were considered as the I-set according to their intermediate nature (Haxpaz & Chothia 1994).

Discussion

The present invention relates to the problem of identifying high order features and organization within the protein sequence space. It is an object of the invention to exhaustively "chart" proteins, preferably all proteins, and to automatically classify them into families. According to the invention, based on pairvvise similarities, statistical regularities in the sequence space are sought, and, more concretely, clusters of protein sequences. According to the invention, in this high order organization most groups consist of sequences with a specific biological function or structure.

A complete charting of the protein space is a daunting task and many difficulties are encountered. One preferably begins from well established statistical measures, in order to identify significant similarities. Great caution and biological expertise are needed to exclude connections which are unacceptable or misleading. The main difficulties stem from chance similarities among sequences and multi-domain-based connections. However, the sheer volume of data makes it necessary to develop automatic methods for such identification.

The present invention addresses these major problems. According to the invention, one starts by creating, for each protein sequence, an exhaustive list of neighboring sequences. These lists take into account the scores of the three major methods for pairwise sequence comparisons. These three types of scores are jointly normalized and the lists are filtered. The link is maintained in the lists only when a significant relationship seems to exist. Then a two-phase clustering algorithm is applied to identify groups of related sequences. There are two pitfalls to avoid here: on the one

hand, it is very easy to follow a very strict rule and generate many small clusters, within each of which, the proteins are very closely related. This approach safely avoids the creation of false connections. However, it adds very little to the understanding of the protein space, since it includes only fairly obvious and well known connections. The other extreme strategy is not useful either: a procedure that declares a connection without sufficient scrutiny, does not miss interesting connections, but generates so many false connections that it becomes impossible to recognize significant relations. In other words, such a permissive method quickly collapses the whole space into a small number of gigantic meaningless clusters. The problem is to find a golden path where non obvious relationships are discovered without 'littering' these valuable connections with too many false connections. The method of the present invention employs an algorithm that can be described as a moderate version of the transitive closure algorithm. At each round of this process statistical information is gained on the relationships among current clusters. This information is then used to merge certain clusters, thus forming the next round of larger, coarser clusters. The algorithm starts from a very conservative classification, and is repeatedly applied, at varying levels of confidence, the input for each stage being the classification output at the previous stage.

Finally, a hierarchical organization of all protein sequences is obtained, strongly correlated with a functional partitioning of all proteins. This data structure reveals interesting relations between and within protein families, and provides a global view ("map") of the space of all proteins.

The classification consists of several thousand clusters, the largest of which contain several hundred members each. It is interesting to assess the effect of slowing down the transitive closure algorithm. Indeed, a straight forward application of the transitive closure (a.k.a. single-linkage clustering) algorithm leads to an avalanche as discussed above (details not shown). Already at confidence level 10^{-20}, most of the space is made up of a small number of very large clusters. This avalanche is caused by chance similarities and chains of domain-based connections that cause unrelated families to merge into few giant clusters. A major ingredient of the new algorithm is

the choice of rules for avoiding such undesirable incorrect connections and for preventing this collapse, and indeed most of them were eliminated.
Some domain-based connections did escape the filters, and caused unrelated clusters to merge. Also, high scoring low-complexity segments which may be biologically meaningless can lead to false connections, and to the formation of nonhomogeneous clusters. Though we did take into account the effect of these segments, not all of them were filtered out. Since the goal was to detect many remote homologies weak filters were used in this case, and many similarities were considered at very low levels of confidence. At the moment, domain based connections can be easily traced manually, by observing the alignments.

It is very difficult or even impossible to properly classify all proteins. The space of proteins has many different facets, all of which should be considered in future, more thorough classification: 3D biological function, biological function, domain content, cellular location, tissue specificity, metabolic pathways, etc. Here a functional partitioning was sought, which is based on sequence analysis. This work differs from previous large scale analyses in several ways.- (i) There is no attempt to identify protein domains or motifs. (ii) No pre-defined groups or other classification are being employed in the analysis of the invention. Moreover, no multiple alignments of the proteins are needed. (iii) The space of all protein sequences is charted not just particular families. (iv) A global organization of all protein sequences is offered.

The algorithm employed in the method of the present invention has turned out well defined groups which are strongly correlated with protein families and subfamilies when compared with the well accepted databases PROSITE and Pfam, the classification performed well for most of the families, though many of them were domain based rather than protein families. Many new clusters in the classification didn't have a match from either PROSITE nor Pfam. The hierarchical organization can indicate the existence of subfamilies within families, and the concept of possibly related clusters exposes distant relationships which reflect functional similarities. These relations offer a basis for sketching local maps near protein families. These connections can be of biological interest, and should be taken into account in the

study of protein families. For example, an overall view of Fig. 17 or Fig. 18 implies that tracing the hierarchical organization of a cluster and/or its rejected merges can provide new information.

Another aspect of the possibly related clusters is that this list can be viewed as a "soft clustering", where the same protein can participate in several different clusters, Proteins that are composed of several domains, some of which shared by proteins from different families, are naturally associated with more than one group. Therefore, though in one embodiment of the invention a multi-domain protein is classified to a single cluster, its multi trait feature is tractable if the relations with the other clusters are examined as well.

References

[Altschul et al. 1990] Altschul, S. F., Carrol, R. J. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

[Altschul 1991] Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219, 555-565.

[Altschul et al. 1994] Altschul, S. F., Boguski, M. S., Gish, W. G. & Wooton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genetics* 6, 119-129.

[Attwood et al. 1998] Attwood, T. K., Beck, M. E., Flower, D. R., Scordis, P. & Selley, J. (1998). The PRINTS protein fingerprint database in its fifth year. *Nucl. Acids Res.* 26, 304-308.

[Bairoch & Boeckman 1992] Bairoch, A. & Boeckman, B. (1992). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* 20, 2019-2022.

[Bairoch et al. 1997] Bairoch A., Bucher P., & Hofmann K. (1997). The PROSITE database, its status in 1997. *Nucl. Acids Res.* 25, 217-221.

[Bockaert 1991] Bockaert, J. (1991). G-proteins and G-protein-coupled receptors:structure, function and interactions. *Curr. Opin. Neurobiol.* 1, 32-42.

[Brenner et al. 1998] Brenner, S.E., Chothia, C. & Hubbard, T. J. P. (1998). *Proc. Natl. Acad. Sci. USA* 95, 6073-6078.

[Chothia 1992] Chothia, C. (1992). One thousand families for the molecular biologist. *Nature* 357, 543-544.

[Cockcroft et al. 1993] Cockcroft, V. B., Ortells, M. O., Thomas, P. & Lunt, G. G. Homologies and disparities of glutamate receptors: a critical analysis. (1993). *Neurochem. Int.* 23, 583-594.

[Compugen] Compugen LTD. http://www.compugen.co.il

[Doolittle 1992] Doolittle, R. F. (1992). Reconstructing history with amino acid sequences. *Protein Sci.* 1, 191-200.

[Doolittle 1998] Doolittle, R. F. (1998). Microbial genomes opened up. *Nature* 392, 339-342.

[Downward 1990] Downward, J. (1990). The ras superfamily of small GTP-binding proteins. *Trends Biochem. Sci.* 15, 469-472.

[Ferran et al. 1994] Ferran, E. A., Pflugfelder, B. & Ferrara P. (1994). Self-Organized Neural Maps of Human Protein Sequences. *Protein Sci.* 3, 507-521.

[Flores et al. 1993] Flores, T. P., Orengo, C. A., Moss, D. & Thoronton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* 2, 1811-1826.

[George et al. 1996] George, D. G., Barker, W. C., Mewes, H. W., Pfeiffer, F. & Tsugita, A. (1996). The PIR-International protein sequence database. *Nucl. Acids. Res.* 24, 17-20.

[Gonnet et al. 1992] Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science* 256, 1443-1445.

[Gracy & Argos 1998] Gracy, J. & Argos, P. (1998). Automated protein sequence database classification. I. Integration of copositional similarity search, local similarity search and multiple sequence alignment. II. Delination of domain boundries from sequence similarity. *Bioinformatics* 14:2, 164-187

[Han & Baker 1995] Han, K. F. & Baker, D. (1995). Recurring local sequence motifs in proteins. *J. Mol. Biol.* 251, 176-187.

[Hanke et al. 1996] Hanke, J., Beckmann, G., Bork, P. & Reich, J. G. (1996). Self-organizing hierarchic networks for pattern recognition in protein sequence. *Protein Sci.* 5, 72-82.

[Harpaz & Chothia 1994] Harpez, Y. & Chothia, C. (1994). Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J. Mol. Biol.* 238, 528-539.

[Harris et al. 1992] Harris, N. L., Hunter, L. & States, D.J. (1992). Mega-classification: Discovering motifs in massive datastreams. In *Proc. of the 10th national conf. on AI*, 837-842, AAAI press/The MIT Press, Menlo park/Cambridge.

[Hilbert et al. 1993] Hilbert, M., Bohm, G. & Jaenicke, R. (1993). Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins* 17, 138-151.

[Henikoff & Henikoff 1991] Henikoff, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucl. Acids Res.* 19, 6565-6572.

[Henikoff & Henikoff 1992] Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* 89, 10915-10919.

[Hobohm and Sander 1995] Hobohm, U. & Sander, C. (1995). A sequence property approach to searching protein database. *J. Mol. Biol.* bf 251, 390-399.

[Holmes et al. 1993] Holmes, K. C., Sander, C. & Valencia, A. (1993). A new ATP-binding fold in actin, hexokinase and Hsc70. *Trends Cell Biol.* 3, 53-59.

[Karlin & Altschul 1990] Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA* 87, 2264-2268.

[Koonin et al. 1996] Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1996). Protein sequence comparison at genome scale. *Methods Enzymol* 266, 295-321.

[Krause & Vingron 1998] Krause, A. & Vingron, M. (1998). A set-theoretic approach to database searching and clustering. Antje Krause and Martin Vingron. *Bioinformatics* 14:5, 430-438.

[Linial et al. 1997] Linial, M., Linial, N., Tishby, N. & Yona, G. (1997). Global self organization of all known protein sequences reveals inherent biological signatures. *J. Mol. Biol.* 268, 539-556.

[Lipman & Pearson 1985] Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity. *Science* 227, 1435-1441.

[Murzin 1993] Murzin, A. G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J.* 12(8), 861-867.

[Murzin et al. 1995] Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.

[Needleman & Wunsch 1970] Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443-453.

[Neuwald et al. 1997] Neuwald, A. F., Liu, J. S., Lipman, D. J. & Lawrence, C. E. (1997). Extracting protein alignment models from the sequence database. *Nucl. Acids Res.* 25(9), 1665-1677.

[Nuoffer & Balch 1994] Nuoffer, C. & Balch, W. (1994). GTPase: multifunctional molecular switches regulating vesicular traffic. *Ann. Rev. Biochem.* 63, 949-990.

[Park et al. 1997] Park J., Teichmann, S. A., Hubbard, T. & Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* 273, 349-354.

[Pearson 1995] Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci.* 4, 1145-1160.

[Pearson 1996] Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol* 266, pp 227-258.

[Pearson 1997] Pearson, W. R. (1997). Identifying distantly related protein sequences. *Comp. App. Biosci.* 13(4), 325-332.

[Pearson 1998] Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* 276, 71-84.

[Pennisi 1997] Pennisi, E. (1997). Microbial genomes come tumbling in. *Science* 277, 1433.

[Pisier 1997] Pisier, G. (1989). The volume of convex bodies and Banach space geometry. Cambridge University Press, Cambridge.

[Sander & Schneider 1991] Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9, 66-68.

[Sheridan & Venkataraghavan 1992] Sheridan, R.P. & Venkataraghavan, R. (1992). A systematic search for protein signature sequences. *Proteins* 14, 16-28.

[Sonnhammer & Kahn 1994] Sonnhammer, E. L. L. & Kahn, D. (1994). Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.* 3, 482-492.

[Sonnhammer et al. 1997] Sonnhammer, E. L., Eddy, S. R., Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28, 405-420.

[Smith & Waterman 1981] Smith, T. F. & Waterman, M. S. (1981). Comparison of Biosequences. *Adv. App. Math.* 2, 482-489 .

[Smith & Xue 1997] Smith, D. K. & Xue, H. (1997). Sequence profiles of immunoglobulin and immunoglobulin-like domains. *J. Mol. Biol.* 274, 530-545.

[Tatusov et al. 1997] Tatusov, R. L., Eugene, V. K. & David, J. L. (1997). A genomic perspective on protein families. *science* 278, 631-637.

[Wang 1996] Wang, Z. (1996). How many fold types of protein are there in nature? *Proteins* 26, 186-191.

[Watanabe & Otsuka 1995] Watanabe, H. & Otsuka, J. (1995). A comprehensive representation of extensive similarity linkage between large numbers of proteins. *Comp. App. Biosci.* 11(2), 159-166.

[Wootton & Federhen 1993] Wootton, J. C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comp. Chem.* 17, 149-163.

[Wu et al. 1992] Wu, C., Whitson, G., Mclarty, J., Ermongkonchai A. & Chang, T. (1992). Protein classification artificial neural system. *Protein Sci.* 1, 667-677.

[van Heel et al. 1991] van Heel, M. (1991). A new family of powerful multivariate statistical sequence analysis techniques. *J. Mol. Biol.* 220, 877-887.